Shaping Biomedicine as an Information Science

Timothy Lenoir

A New Biology for the Information Age

Sometime in the mid-1960s biology became an information science. While François Jacob and Jacques Monod's work on the genetic code is usually credited with propelling biology into the Information Age, in this essay I explore the transformation of biology by what have become essential tools to the practicing biochemist and molecular biologist: namely, the contributions of information technology. About the same time as Jacob and Monod's work, developments in computer architectures and algorithms for generating models of chemical structures and simulations of chemical interactions were created that allowed computational experiments to interact with and draw together theory and laboratory experiment in completely novel ways. The new computational science linked with visualization has had a major impact in the fields of biochemistry, molecular dynamics, and molecular pharmacology (Friedhoff & Benzon, 1989; Panel on Information Technology and the Conduct of Research, 1989; McCormick, DeFanti, & Brown, 1987, p. A-1; Hall, 1995). By "computational science" I mean the use of computers in science disciplines like these as distinct from computer science (McCormick, DeFanti, & Brown, p. 11). The sciences of visualization are defined by McCormick, DeFanti, and Brown as follows:

> Images and signals may be captured from cameras or sensors, transformed by image processing, and presented pictorially on hard or soft copy output. Abstractions of these visual representations can be transformed by computer vision to create symbolic representations in the form of symbols and structures. Using computer graphics, symbols or structures can be synthesized into visual representations. (P. A-1)

Computational approaches have substantially transformed and extended the domain of theorizing in these areas in ways unavailable to older, non–computer-based forms of theorizing.

But other information technologies have also proved crucial to bringing about this change. In the 1970s through the 1990s, armed with such new tools of molecular biology as cloning, restriction enzymes, protein sequencing, and gene product amplification, biologists were awash in a sea of new data. They deposited this data in large and growing electronic databases of genetic maps, atomic coordinates for chemical and protein structures, and protein sequences. These developments in technique and instrumentation launched biology onto the path of becoming a data-bound science, a "science" in which *all* the data of a domain—such as a genome—are available before the laws of the domain are understood. Biologists have coped with this data explosion by turning to information science: applying artificial intelligence and expert systems and developing search tools to identify structures and patterns in their data.

The aim of this paper is to explore early developments in the introduction of computer modeling tools from artificial intelligence (AI) and expert systems into biochemistry in the 1960s and 1970s, and the introduction of informatics techniques for searching databases and extracting biological function and structure in the emerging field of genomics during the 1980s and 1990s. I have two purposes in this line of inquiry. First I want to suggest that by introducing tools of information science biologists have sought to make biology a unified theoretical science with predictive powers analogous to other theoretical disciplines. But I want also to suggest that along with this highly heterogeneous and hybrid form of computer-based experimentation

and theorizing has come a different conception of theorizing itself: one based on models of information-processing and best captured by the phrase "knowledge engineering" developed within the AI community. My second concern is to contribute to recent discussions on the transformation of biology into an information science. Lily Kay, Evelyn Fox Keller, Donna Haraway, and Richard Doyle have explored the role of metaphor, disciplinary politics, economics, and culture in shaping the context in which the language of "DNA code," "genetic information," "text," and "transcription" have been inserted into biological discourse, often in the face of resistance from some of the principal actors themselves (Doyle, 1997). I am more interested than these authors in software and the computational regimes that it enables. Elaborating on the theme of "tools to theory," recently espoused in science and technology studies, I am interested in exploring the role of the computational medium itself in shaping biology as an information science. But a further crucial stimulation to the takeoff of bioinformatics, of course, was provided by hardware and networking developments underwritten by the NIH and NSF (Hughes, 1999).

Computers and Biochemistry: Molecular Modeling

The National Institutes of Health have been active at every stage in making biology an information science. NIH support was crucial to the explosive take-off of computational chemistry and the promotion of computer-based visualization technologies in the mid-1960s. The agency sponsored a conference at UCLA in 1966 on "Image Processing in Biological Science." The NIH's Bruce Waxman, co-chair of the meeting, set out the NIH agenda for computer visualization by sharply criticizing the notion of mere "image processing" as the direction that should be pursued in computer-enhanced vision research. The goal of computer-assisted "vision," he asserted, was not to replicate relatively low-order motor and perceptual capabilities even at rapid speeds. "I have wondered whether the notion of image processing is itself restrictive; it may connote the reduction of and analysis of 'natural' observations but exclude from consideration two- or three-dimensional data which are abstractions of phenomena rather than the phenomena themselves" (Ramsey, 1968, pp. xiii–xiv). Waxman suggested "pattern recognition" as the subject that they should really pursue—and in particular where the object was what he termed "non-natural." In general, Waxman asserted, by its capacity to quantize massive data sets automatically, the computer, linked with pattern-recognition methods of imaging the non-natural, would permit the development of stochastically based biological theory.

Waxman's comments point to one of the important and explicit goals of the NIH and other funding agencies: to mathematize biology. That biology should follow in the footsteps of physics had been the centerpiece of a reductionist program since at least the middle of the nineteenth century. But the development of molecular biology in the 1950s and 1960s encouraged the notion that a fully quantitative theoretical biology was on the horizon. The computer was to be the motor for this change. Analogies were drawn between highly mathematized and experimentally based Big Physics and the anticipated "Big Biology." As Lee B. Lusted, the chairman of the Advisory Committee to the National Research Council on Electronic Computers in Biology and Medicine argued, because of the high cost of computer facilities for conducting biological research, computer facilities would be to biology what SLAC (the Stanford Linear Accelerator) and the Brookhaven National Laboratory were to physics (Ledley, 1965, pp. ix–x). Robert Ledley, then affiliated with the Division of Medical Sciences, National Research Council, and author of the volume expressed the committee's interest in fostering computing and insisted that biology was on the threshold of a new era. New emphasis on quantitative work and experiment was changing the conception of the biologist: the view of the biologist as an individual scientist, personally carrying through each step of his investigation and his data-reduction processes, was rapidly broadened to include the biologist as a part of an intricate organizational chart that partitions scientific, technical, and administrative responsibilities. In the new organization of biological work, modeled on the physicists' work at large national labs, the talents and knowledge of the biologist "must be augmented by those of the engineer, the mathematical analyst, and the professional computer programmer" (Ledley, 1965, p. xi).

At the UCLA meeting Bruce Waxman held up as a model the work on three-dimensional representations of protein molecules carried out by Cyrus Levinthal. Levinthal worked with the facilities of MIT's Project on Mathematics and Computation (MAC), one of the first centers in the development of graphics. Levinthal's project was an experiment in computer time-sharing linking biologists, engineers, and mathematicians in the construction of Big Biology. Levinthal's work at MIT illustrates the role of computer visualization as a condition for theory development in molecular biology and biochemistry.

Since the work of Linus Pauling and Robert Corey

on the α-helical structure of most protein molecules in 1953, models have played a substantial role in biochemistry. Watson and Crick's construction of the double helix model for DNA depended crucially upon the construction of a physical model. Subsequently, work in the field of protein biology has demonstrated that the functional properties of a molecule depend not only on the interlinkage of its chemical constituents but also on the way in which the molecule is configured and folded in three dimensions. Much of biochemistry has focused on understanding the relationship between biological function and molecular conformational structure.

A milestone in the making of physical models (in three dimensions) of molecules was John Kendrew's construction of myoglobin. The power of models in investigations of biomolecular structure was evident from work such as this, but such tools had limitations as well. Kendrew's model, for instance, was the first successful attempt to build a physical model into a Fourier map of a molecule's electron densities derived from X-ray crystallographic sources. As a code for electron density, clips of different colors were put at the proper vertical positions on a forest of steel rods. A brass wire model of the alpha helices and beta sheets that make up the molecule was then built in among the rods. Mechanical interference made it difficult to adjust the structure, and the model was hard to see because of the large number of supporting rods. The model incorporated both too little and too much: too little, in that the basic shape of the molecule was not represented; too much, in that the forest of rods made it difficult to see the three-dimensional folding of the molecule (even though bond connectivity was represented). Perhaps the greatest drawback was the model's size: It filled a large room. The answer to these problems was computer representation. For an early stereogram of myoglobin constructed by computer on the basis of Kendrew's data, see Watson (1969). It was obvious that such three-dimensional representations would only become really useful when it was possible to manipulate them at will. Proponents of computer graphics argued that this flexibility is exactly what computer representations of molecular structure would allow. Cyrus Levinthal first illustrated these methods in 1965.

Levinthal reasoned that since protein chains are formed by linking molecules of a single class, amino acids, it should be relatively easy to specify the linkage process in a form mathematically suitable for a digital computer (Levinthal, 1966). Initially the computer model considers the molecule as a set of rigid groups of constant geometry linked by single bonds around which rotation is possible. Program input consists of a set of coordinates consistent with the molecular stereochemistry as given in data from X-ray crystallographic studies. Several constraints delimit stable configurations among numerous possibilities resulting from combinations of linkages among the twenty different amino acids. These include bond angles, bond lengths, van der Waals radii for each species of atom, and the planar configuration of the peptide bond.

Molecular biologists, particularly the biophysicists among them, were motivated to build a unified theory, and the process of writing a computer program that could simulate protein structure would assist in this goal by providing a framework of mental and physical discipline from which would emerge a fully mathematized theoretical biology. In such non-mathematized disciplines as biology, the language of the computer program would serve as the language of science (Oettinger, 1966, p. 161). But there was a hitch: In an ideal world dominated by a powerful central theory, one would like, for example, to use the inputs of xyz coordinates of the atoms, types of bond, and so forth, to calculate the pairwise interaction of atoms in the amino acid chain, predict the conformation of the protein molecule, and check this against its corresponding X-ray crystallographic image. As described, however, the parameters used as input in the computer program do not provide much limitation on the number of molecular conformations. Other sorts of input are needed to filter among the myriad possible structures. Perhaps the most important of these is energy minimization. In explaining how the thousands of atoms in a large protein molecule interact with one another to produce a stable conformation, one hypothesizes that, like water running downhill, the molecular string will fold to reach a lowest energy level. To carry out this sort of minimization would entail calculating the interactions of all pairs of active structures in the chain, minimizing the energy corresponding to these interactions over all possible configurations, and then displaying the resulting molecular picture. Unfortunately, this objective could not be achieved, as Levinthal noted, because a formula describing such interactions could not, given the state of molecular biological theory in 1965, even be stated, let alone be manipulated with a finite amount of labor. In Levinthal's words:

> The principal problem, therefore, is precisely how to provide correct values for the variable angles. . . . I should emphasize the magnitude of the problem that remains even after one has gone as far as possible in using chemical constraints to reduce the number of variables from

> several thousand to a few hundred. . . . I therefore decided to develop programs that would make use of a man-computer combination to do a kind of model-building that neither a man nor a computer could accomplish alone. This approach implies that one must be able to obtain information from the computer and introduce changes in the way the program is running in a span of time that is appropriate to human operation. This in turn suggests that the output of the computer must be presented not in numbers but in visual form. (Levinthal, 1966, pp. 48–49)

In Levinthal's view, visualization generated in real-time interaction between human and machine can assist theory construction. The computer becomes in effect both a microscope for examining molecules as well as a laboratory for quantitative experiment. Levinthal's program, CHEMGRAF, could be programmed with sufficient structural information as input from physical and chemical theory to produce a trial molecular configuration as graphical output. A subsystem called SOLVE then packed the trial structure by determining the local minimum energy configuration due to non-bonded interactive forces. A subroutine of this program, called ENERGY, calculated the torque vector caused by the atomic pair interactions on rotatable bond angles. An additional procedure for determining the conformation of the model structure was "cubing." This procedure searched for nearest neighbors of an atom in the center of a $3 \times 3 \times 3$ cube and reported whether any atoms were in the twenty-six adjacent cubes. The program checked for atom pairs in the same or adjacent cubes and for atoms within a specified distance. It maintained a list of pairs that were, for instance, in contact violation, while another routine calculated energy contribution of the pair to the molecule. The cubing program rejected as early as possible all those atom pairs where the interatomic distance was too great to be of more than negligible contribution, and it enabled more efficient use of computer time.

Levinthal emphasized that interactivity was a crucial component of CHEMGRAF. Built into his system was the requirement of observing the result of the calculations interactively so that one could halt the minimization process at any step, either to terminate it completely or to alter the conformation and then resume it (Katz & Levinthal, 1972). Levinthal noted that often, as the analytical procedures were grinding on, a molecule would be trapped in an unfavorable conformation or in a local minimum and the problem would be obscure until the conformation could be viewed three-dimensionally. CHEMGRAF enabled the investigator to assist in generating the local minimization of energy for a subsection of the molecule through three different types of user-guided empirical manipulation of structure: "close," "glide," and "revolve." These manipulations in effect introduced external "pseudo-energy" terms into the computation that pulled the structure in various ways (Levinthal, Barry, Ward, & Zwick, 1968). Atoms could be rotated out of the way by direct command and a new starting conformation chosen from which to continue the minimization procedure. By pulling individual atoms to specific locations indicated by experimental data from X-ray diffraction studies, a fit between X-ray crystallographic data and the computer model of a specific protein, such as myoglobin, could ultimately be achieved. With the model in hand of the target molecule, such as myoglobin, one could then proceed to investigate the various energy terms involved in holding the protein molecule together. Thus, the goal of this interactive effort involving human and machine was eventually to generate a theoretical formulation for the lowest energy state of a protein molecule, to predict its structure, and to have that prediction confirmed by X-ray crystallographic images (Hall, 1995).

The enormous number of redundant trial calculations involved in Levinthal's work hints at the desirability of combining an expert system with a visualization system. E. J. Corey and W. Todd Wipke, working nearby at Harvard, took this next step. (Space limitations prevent me from discussing their work here.) In developing their work, Wipke and Corey drew upon a prototype expert system at Stanford called DENDRAL, the result of a collaboration at Stanford among computer scientist Edward Feigenbaum, biologist Joshua Lederberg, and organic chemist Carl Djerassi, working on another of the NIH initiatives to bring computers directly into the laboratory. The Stanford project, called DENDRAL, was an early effort in the field of what Feigenbaum and his mentors Herbert Simon and Marvin Minsky termed "knowledge engineering." In effect, it attempted to put the human inside the machine.

DENDRAL: The AI Approach at Stanford

DENDRAL aimed at emulating an organic chemist operating in the harsh environment of Mars (Lederberg, n.d.; Lederberg, Sutherland, Buchanan, & Feigenbaum, 1969). The ultimate goal was to create an automated laboratory as part of the Viking mission planned to land a mobile instrument pod on Mars in 1975. Given the

mass spectrum of an unknown compound, the specific goal was to determine the structure of the compound. To accomplish this, DENDRAL would analyze the data, generate a list of plausible candidate structures, predict the mass spectra of those structures from the theory of mass spectrometry, and select as a hypothesis the structure whose spectrum most closely matched the data.

A key part of this program was the representation of chemical structure in terms of topological graph theory. Chemical graphs were the visual "language" to augment the theoretical and practical knowledge of the chemist with the calculating power of the computer. This part of the effort was contributed by Lederberg, the winner of the 1958 Nobel Prize in medicine or physiology, for his work on genetic exchange in bacteria, who had been interested in the introduction of information concepts into biology for most of his professional life. Self-described as a man with a Leibnizian dream of a universal calculus for the alphabet of human thought, Lederberg's interest in mass spectrometry and topological mapping of molecules was in part driven by the dream of mathematizing biology, starting with organic chemistry. The structures of organic molecules are bewilderingly complex, and the "theory" of organic chemistry does not have an elegant axiomatic structure analogous, say, to Newtonian mechanics, even though it is sprinkled with lots of theory derived from quantum mechanics and thermodynamics. Lederberg felt that a first step toward such a quantitative, predictive theory would be a rational systematization of organic chemistry. Trampling upon a purist's notion of theory, Lederberg thought that computers were the royal road to mathematization in chemistry:

> Could not the computer be of great assistance in the elaboration of novel and valid theories? I can dream of machines that would not only execute experiments in physical and chemical biology but also help design them, subject to the managerial control and ultimate wisdom of their human programmer. (Lederberg, 1969)

Mass spectrometry, the area upon which Feigenbaum and Lederberg concentrated with Carl Djerassi, was a particularly appropriate challenge. It differed in at least one crucial aspect from the molecular modeling of proteins I have considered above. Whereas in those areas a well-understood theory, such as the quantum mechanical theory of the atomic bond, was the basis for developing the computer program to examine effects in large calculations, there was no theory of mass spectrometry that could be transferred to the program from a textbook (Lederberg, Sutherland, Buchanan, & Feigenbaum, 1969). The field has bits of theory to draw upon, but it has developed mainly by following rules of thumb, which are united in the form of the chemist-expert. The field thrives on tacit knowledge. The following excerpt from a memo by Feigenbaum written after his first meetings with Lederberg on the DENDRAL project provides a vivid sense of the objective and the problems faced:

> The main assumption we are operating under is that the required information is buried in chemists' brains if only we can extract it. Therefore, the initiative for the interaction must come from the system not the chemist, while allowing the chemist the flexibility to supply additional information and to modify the question sequence or content of the system. . . . What we want to design then is a question asking system [that] will gather rules about the feasibility of the chemical molecules and their subgraphs being displayed. ("Second Cut," n.d.)

In short, Feigenbaum sought to emulate a gifted chemist with the computer. That chemist was Carl Djerassi, nicknamed "El Supremo" by his graduate and postdoctoral students. Djerassi's astonishing achievements as a mass spectrometrist relied on his abilities to feel his way through the process without the aid of a complete theory, relying rather on experience, tacit knowledge, hunches, and rules of thumb. In interviews Feigenbaum elicited this kind of information from Djerassi, in a process that heightened awareness of the structure of the field for both participants. The process of involving a computer in chemical research in this way organized a variety of kinds of information, which constituted a crucial step toward theory.

A Paradigm Shift in Biology

Thus far I have been considering efforts to predict structure from physical principles as the first path through which computer science and computer-based information technology began to reshape biology. The Holy Grail of biology has always been the construction of a mathematized theoretical biology, and for most molecular biologists the journey there has been directed by the notion that the information for the three-dimensional folding and structure of proteins is uniquely contained in the linear sequence of their amino acids (Anfinsen, 1973). As we have seen, the molecular dynamics approach assumed that if all the forces between atoms in a molecule, including bond energies and electrostatic attraction and repulsion, are known, then it is possible to calculate the three-dimensional arrangement of atoms

that requires the least energy. Christian B. Anfinsen (1973) discussed the work for which he was awarded the Nobel Prize in chemistry in 1972:

> This hypothesis (the "thermodynamic hypothesis") states that the three-dimensional structure of a native protein in its normal physiological milieu . . . is the one in which the Gibbs free energy of the whole system is lowest; that is, that the totality of interatomic interactions and hence by the amino acid sequence, in a given environment. (P. 223)

Because this method requires intensive computer calculations, shortcuts have been developed that combine computer-intensive molecular dynamics computations, artificial intelligence, and interactive computer graphics in deriving protein structure directly from chemical structure.

While theoretically elegant, the determination of protein structure from chemical and dynamical principles has been hobbled with difficulties. In the abstract, analysis of physical data generated from protein crystals, such as X-ray and nuclear magnetic resonance data, should offer rigorous ways to connect primary amino acid sequences to three-dimensional structure. But the problems of acquiring good crystals and the difficulty of getting NMR data of sufficient resolution are impediments to this approach. Moreover, while quantum mechanics provides a solution to the protein-folding problem in theory, the computational task of predicting structure from first principles for large protein molecules containing many thousands of atoms has proved impractical. Furthermore, unless it is possible to grow large, well-ordered crystals of a given protein, X-ray structure determination is not an option. The development of methods of structure determination by high-resolution two-dimensional NMR has alleviated this situation somewhat, but this technique is also costly and time-consuming, requiring large amounts of protein of high solubility, and is severely limited by protein size. These difficulties have contributed to the slow rate of progress in registering atomic coordinates of macromolecules.

An indicator of the difficulty of pursuing this approach alone is suggested by the relatively slow growth of databanks of atomic coordinates for proteins. The Protein Data Bank (PDB) was established in 1971 as a computer-based archival resource for macromolecular structures. The purpose of the PDB was to collect, standardize, and distribute atomic coordinates and other data from crystallographic studies. In 1977 the PDB listed atomic coordinates for forty-seven macromolecules (Bernstein et al., 1977). In 1987 that number began to increase rapidly at a rate of about 10 percent per year because of the development of area detectors and widespread use of synchrotron radiation; by April 1990 atomic coordinate entries existed for 535 macromolecules. Commenting on the state of the art in 1990, Holbrook and colleagues (1993) noted that crystal determination could require one or more man-years. Currently (1999), the PDB's Biological Macromolecule Crystallization Database (BMCD) contains entries for 2,526 biological macromolecules for which diffraction quality crystals have been obtained. These include proteins, protein:protein complexes, nucleic acid, nucleic acid:nucleic acid complexes, protein:nucleic acid complexes, and viruses.[1]

While structure determination was moving at a snail's pace, beginning in the 1970s, another stream of work contributed to the transformation of biology into an information science. The development of restriction enzymes, recombinant DNA techniques, gene cloning techniques, and polymerase chain reactions (PCRs) resulted in a flood of data on DNA, RNA, and protein sequences. Indeed more than 140,000 genes were cloned and sequenced in the twenty years from 1974 to 1994, of which more than 20 percent were human genes (Brutlag, 1994, p. 159). By the early 1990s, well before the beginning of the Human Genome Initiative, the NIH GenBank database (release 70) contained more than 74,000 sequences, while the Swiss Protein database (Swiss-Prot) included nearly 23,000 sequences. Protein databases were doubling in size every twelve months, and some were predicting that by the year 2000 ten million base pairs a day would be sequenced as a result of the technological impact of the Human Genome Initiative. Such an explosion of data encouraged the development of a second approach to determining the function and structure of protein sequences: namely, prediction from sequence data alone. This "bioinformatics" approach identifies the function and structure of unknown proteins by applying search algorithms to existing protein libraries in order to determine sequence similarity, percentages of matching residues, and the statistical significance of each database sequence.

A key project illustrating the ways in which com-

[1] Biological Macromolecule Crystallization Database and the NASA Archive for Protein Crystal Growth Data (version 2.00) are located on the Web at http://www.bmcd.nist.gov:8080/bmcd/bmcd.html.

puter science and molecular biology began to merge in the formation of bioinformatics was the MOLGEN project at Stanford and events related to the formation and subsequent development of BIONET. MOLGEN was a continuation of the projects in artificial intelligence and knowledge engineering begun at Stanford with DENDRAL. MOLGEN was started in 1975 as a project in the Heuristic Programming Project with Edward Feigenbaum as principal investigator directing the thesis projects of Mark Stefik and Peter Friedland (Feigenbaum & Martin, 1977). The aim of MOLGEN was to model the experimental design activity of scientists in molecular genetics (Friedland, 1979). Before an experimentalist sets out to achieve some goal, he produces a working outline of the experiment, guiding each step of the process. The central idea of MOLGEN was based on the observation that scientists rarely plan from scratch in designing a new experiment. Instead, they find a skeletal plan, an overall design that has worked for a related or more abstract problem, and then adapt it to the particular experimental context. Like DENDRAL, this approach is heavily dependent upon large amounts of domain-specific knowledge in the field of molecular biology and even more upon good heuristics for choosing among alternative implementations.

MOLGEN's designers chose molecular biology as appropriate for the application of artificial intelligence because the techniques and instrumentation generated in the 1970s seemed ripe for automation. The advent of rapid DNA cloning and sequencing methods had had an explosive effect on the amount of data that could be most readily represented and analyzed by a computer. Moreover, it appeared that very soon progress in analyzing information in DNA sequences would be limited by the lack of an appropriate combination of search and statistical tools. MOLGEN was intended to apply rules to detect profitable directions for analysis and to reject unpromising ones (Feigenbaum et al., 1980).

Peter Friedland was responsible for constructing the knowledge-base component of MOLGEN. Though not himself a molecular biologist, he made a major contribution to the field by assembling the rules and techniques of molecular biology into an interactive, computerized system of analytical programs. Friedland worked with Stanford molecular biologists Douglas Brutlag, Laurence Kedes, John Sninsky, and Rosalind Grymes, who provided expert knowledge on enzymatic methods, nucleic acid structures, detection methods, and pointers to key references in all areas of molecular biology. Along with providing an effective encyclopedia of information about technique selection in planning a laboratory experiment, the knowledge base contained a number of tools for automated sequence analysis. Brutlag, Kedes, Sninsky, and Grymes were interested in having a battery of automated tools for sequence analysis, and they contracted with Friedland and Stefik—both gifted computer program designers—to build these tools in exchange for contributing their expert knowledge to the project (Douglas Brutlag, personal communication; Peter Friedland, personal communication). (In 1987, after his work on MOLGEN and at IntelliGenetics [discussed below], Friedland went on to become chief scientist at the NASA-Ames Laboratory for Artificial Intelligence.)

This collaboration of computer scientists and molecular biologists helped move biology along the road to becoming an information science. Among the programs Friedland and Stefik created for MOLGEN was SEQ, an interactive self-documenting program for nucleic acid sequence analysis, which had thirteen different procedures with over twenty-five different subprocedures, many of which could be invoked simultaneously to provide various analytical methods for any sequence of interest. SEQ brought together in a single program methods for primary sequence analysis described in the literature by L. J. Korn and colleagues, R. Staden, and numerous others (Korn, Queen, & Wegman, 1977); Staden, 1977; Staden, 1978; Staden, 1979). SEQ also performed homology searches on DNA sequences and specified the degree of homology, and conducted dyad symmetry (inverted repeats) searches (Friedland, Brutlag, Clayton, & Kedes, 1982). Another feature of SEQ was its ability to prepare restriction maps with the names and locations of the restriction sites marked on the nucleotide sequence. In addition it had a facility for calculating the length of DNA fragments from restriction digests of any known sequence. Another program in the MOLGEN suite was GA1 (later called MAP). Constructed by Stefik, GA1 was an artificial intelligence program that allowed the generation of restriction enzyme maps of DNA structures from segmentation data (Stefik, 1977). It would construct and evaluate all logical alternative models that fit the data and rank them in relative order of fit. A further program in MOLGEN was SAFE, which aided in enzyme selection for gene excision. SAFE took amino acid sequence data and predicted the restriction enzymes guaranteed not to cut within the gene itself.

In its first phase of development (1977–1980) MOLGEN consisted of the programs described above

and a knowledge base containing information on about three hundred laboratory methods and thirty strategies for using them. It also contained the best currently available data on about forty common phages, plasmids, genes, and other known nucleic acid structures. The second phase of development beginning in 1980 scaled up the analytical tools and the knowledge base. Perhaps the most significant aspect of the second phase was making MOLGEN available to the scientific community at large on the Stanford University Medical Experimental national computer resource, SUMEX-AIM. SUMEX-AIM, supported by the Biotechnology Resources Program at NIH since 1974, had been home to DENDRAL and several other programs. The new experimental resource on SUMEX, comprising the MOLGEN programs and access to all major genetic databases, was called GENET. In February 1980 GENET was made available to a carefully limited community of users (Rindfleisch, Friedland, & Clayton, 1981).

MOLGEN and GENET were immediate successes with the molecular biology community. In their first few months of operation in 1980 more than two hundred labs (with several users in each of those labs) accessed the system. By 1 November 1982 more than three hundred labs on the system around the clock accessed the system from a hundred institutions (Douglas Brutlag, personal communication; NIH Special Study Section, 1983; Lewin, 1984). Traffic on the site was so heavy that restrictions had to be implemented and plans for expansion considered. In addition to the academic users a number of biotech firms, such as Monsanto, Genentech, Cetus, and Chiron, used the system heavily. Feigenbaum, principal investigator in charge of the SUMEX resource, and Thomas Rindfleisch, facility manager, decided to exclude commercial users in order to ensure that the academic community had unrestricted access to the SUMEX computer and to answer the NIH's concern that commercial users gain unfair access to the resource (Maxam to GENET community, 1982).

To provide commercial users with their own unrestricted access to the GENET and MOLGEN programs, Brutlag, Feigenbaum, Friedland, and Kedes formed a company, IntelliGenetics, which would offer the suite of MOLGEN software for sale or rental to the emerging biotechnology industry. With 125 research labs doing recombinant DNA research in the United States alone and a number of new genetic engineering firms starting up, opportunities looked outstanding. No one was currently supplying software in this rapidly growing genetic engineering marketplace. With their exclusive licensing arrangement with Stanford for the MOLGEN software, IntelliGenetics was poised to lead a huge growth area. The business plan expressed well the excellent position of the company:

> A major key to the success of IntelliGenetics will be the fact that the recombinant DNA research revolution is so recent. While every potential customer is well capitalized, few have the manpower they say they need; this year several firms are hiring 50 molecular geneticist Ph.D.s, and one company speaks of 1000 within five years. These firms require computerized assistance—for the storage and analysis of very large amounts of DNA sequence information which is growing at an exponential rate—and will continue to do so for the foreseeable future (10 years). Access to this information and the ability to perform rapid and efficient pattern recognition among these sequences is currently being demanded by most of the firms involved in recombinant DNA research.
>
> The programs offered by IntelliGenetics will enable the researchers to perform tasks that are: 1) virtually impossible to perform with hand calculations, and 2) extremely time-consuming and prone to human error. *In other words, IntelliGenetics offers researcher productivity improvement to an industry with expanding demand for more researchers which is experiencing a severe supply shortage* [emphasis in original]. ("Business plan for IntelliGenetics," 1981; Friedland to Reimers, 1984)[2]

The resource that IntelliGenetics eventually offered to commercial users was BIONET. Like GENET, BIONET combined all databases of DNA sequences with programs to aid in their analysis in one computer site.

Prior to the startup of BIONET and contemporaneous with GENET, other resources for DNA sequences were developed. Several researchers were making their databases available. Under the auspices of the National Biomedical Research Foundation, Margaret Dayhoff had created a database of DNA sequences and some software for sequence analysis that was marketed commercially. Walter Goad, a physicist at Los Alamos National Laboratory, collected DNA sequences from the published literature and made them freely available to researchers. But by the late 1970s the number of bases sequenced was already approaching three million and expected to

[2] Details of the software licensing arrangement and the revenues generated are discussed in a letter to Niels Reimers, at the Stanford Office of Technology Licensing, on the occasion of renegotiating the terms.

double soon. Some form of easy communication between labs for effective data handling was considered a major priority in the biological community. While experiments were going on with GENET, a number of nationally prominent molecular biologists had been pressing to start an NIH-sponsored central repository for DNA sequences. In 1979, at Rockefeller University, Joshua Lederberg organized an early meeting with such an agenda. The proposed NIH initiative was originally supposed to be coordinated with a similar effort at the European Molecular Biology Laboratory (EMBL) in Heidelberg, but the Europeans became dissatisfied with the lack of progress on the American end and decided to go ahead with their own databank. EMBL announced the availability of its Nucleotide Sequence Data Library in April 1982, several months before the American project was funded. Finally, in August 1982, the NIH awarded a contract for $3 million over five years to the Boston-based firm of Bolt, Berenek, and Newman (BB&N) to set up the national database known as GenBank in collaboration with Los Alamos National Laboratory. IntelliGenetics submitted an unsuccessful bid for that contract.

The discussions leading up to GenBank included consideration of funding a more ambitious databank, known as "Project 2," which was to provide a national center for the computer analysis of DNA sequences. Budget cuts forced the NIH to abandon that scheme (Lewin, 1984). However, officials there returned to it the following year, thanks to the persistence of IntelliGenetics representatives. Although GenBank launched a formal national DNA sequence collection effort, the need for computational facilities voiced by molecular biologists was still left unanswered. In September 1983, after a review process that took over a year, the NIH division of research resources awarded IntelliGenetics a $5.6 million five-year contract to establish BIONET (Lewin, 1984). The contract, the largest award of its kind by the NIH to a for-profit organization (p. 1380), started on 1 March 1984 and ended on 27 February 1989.

BIONET first became available to the research community in November 1984. The fee for use was $400 per year per laboratory and remained at that level throughout its first five years. BIONET's use grew impressively. Initially the IntelliGenetics team set the target for user subscriptions at 250 labs. However, in March 1985, the annual report for the first year's activities of BIONET listed 350 labs with nearly 1,132 users. By August 1985 that number had increased dramatically to 450 labs and 1,500 users (Minutes of the meeting, 1985). In April 1986, for example, BIONET had 464 laboratories comprising 1,589 users. By October 1986 the numbers were 495 labs and 1,716 users (BIONET users status, 1986). By 1989, 900 laboratories in the United States, Canada, Europe, and Japan (comprising about 2,800 researchers) subscribed to BIONET, and 20 to 40 new laboratories joined each month (Huberman, 1989).

BIONET was intended to establish a national computer resource for molecular biology satisfying three goals, which it fulfilled to varying degrees. A first goal was to provide a way for academic biologists to obtain access to computational tools to facilitate research relating to nucleic acids and possibly proteins. In addition to giving researchers ready access to national databases on DNA and protein sequences, BIONET would provide a library of sophisticated software for sequence searching, matching, and manipulation. A second goal was to provide a mechanism to facilitate research into improving such tools. The BIONET contract provided research and development support of further software, both in-house research by IntelliGenetics scientists and through collaborative ventures with outside researchers. A third goal of BIONET was to enhance scientific productivity through electronic communications.

The stimulation of collaborative work through electronic communication was perhaps the most impressive achievement of BIONET. BIONET was much more than the Stanford GENET plus the MOLGEN–IntelliGenetics suite of software. Whereas GENET with its pair of ports could accommodate only two users at any one time, BIONET had twenty-two ports providing an estimated annual thirty thousand connect hours (Friedland, 1984; Smith, Brutlag, Friedland, & Kedes, 1986). All subscribers to BIONET were provided with e-mail accounts. For most molecular biologists this was something entirely new, since most university labs were just beginning to be connected with regular e-mail service. At least twenty different bulletin boards on numerous topics were supported by BIONET. In an effort to change the culture of molecular biologists by accustoming them to the use of electronic communications and more collaborative work, BIONET users were required to join one of the bulletin board groups.

BIONET subscribers had access to the latest versions of the most important databases for molecular biology. Large databases available at BIONET were GenBank, the National Institutes of Health DNA sequence library; EMBL, the European Molecular Biology

Laboratory nucleotide sequence library; NBRF-PIR, the National Biomedical Research Foundation's protein sequence database, which is part of the Protein Identification Resource [PIRI] supported by NIH's Division of Research Resources; SWISS-PROT, a protein sequence database founded by Amos Bairoch of the University of Geneva and subsequently managed and distributed by the European Molecular Biology Laboratory; VectorBank, IntelliGenetics' database of cloning vector restriction maps and sequences; Restriction Enzyme Library, a complete list of restriction enzymes and cutting sites provided by Richard Roberts at Cold Spring Harbor; and Keybank, IntelliGenetics' collection of predefined patterns or "keys" for database searching. Several smaller databases were also available, including a directory of molecular biology databases, a collection of literature references to sequence analysis papers, and a complete set of detailed protocols for use in a molecular biological laboratory (especially for *Escherichia coli* and yeast work) (IntelliGenetics, 1987, p. 23).

Perhaps the most important contribution made by BIONET to establishing molecular biology as an information science did not materialize until the period of the second contract for GenBank. As described above, BB&N was awarded the first five-year contract to manage GenBank. The contract was up for renewal in 1987, and on the basis of its track record in managing BIONET, IntelliGenetics submitted a proposal to manage GenBank. GenBank users had become dissatisfied with the serious delay in sequence data publication. GenBank was two years behind in disseminating sequence data it had received (Douglas Brutlag, personal communication, 19 June 1999). At a meeting in Los Alamos in 1986, Walter Goad noted that GenBank had twelve million base pairs. Other sequence collections available to researchers contained fourteen to fifteen million base pairs, so that GenBank was at least 14 to 20 percent out of date (Boswell, 1987). Concerned that researchers would turn to other, more up-to-date data sources, the NIH listed encouraging use as one of the issues they wanted IntelliGenetics to address in their proposal to manage GenBank (Duke, 1987).

IntelliGenetics proposed to solve this problem by automating the submission of gene and protein sequences. The standard method up to that time required an employee at GenBank to search the published scientific literature laboriously for sequence data, rekey these into a GenBank standard electronic format, and check them for accuracy. IntelliGenetics would automate the submission procedure with an online submission program, XGENPUB (later called "AUTHORIN").

In fact, IntelliGenetics was already progressing toward automating all levels of sequence entry and (as much as possible) analysis. As early as 1986 IntelliGenetics included SEQIN in PC/GENE, its commercial software package designed for microcomputers. SEQIN was designed for entering and editing nucleic acid sequences, and it already had the functionality needed to deposit sequences with GenBank or EMBL electronically ("PC/Gene," 1986). Transferring this program to the mainframe was a straightforward move. Indeed the online entry of original sequence data was already a feature of BIONET, since large numbers of researchers were using the IntelliGenetics GEL program on the BIONET computer. GEL was a program that accepted and analyzed data produced by all the popular sequencing methods. It provided comprehensive record-keeping and analysis for an entire sequencing project from start to finish. The final product of the GEL program was a sequence file suitable for analysis by other programs, such as SEQ.XGENPUB, extended to this capability by allowing the scientist to annotate a sequence according to the standard GenBank format and mail the sequence and its annotation to GenBank electronically. The interface was a forms-oriented display editor that would automatically insert the sequence in the appropriate place in the form by copying the sequence from a designated file on the BIONET computer. When completed, it could be forwarded to the GenBank computer at Los Alamos; the National Institutes of Health DNA sequence library, EMBL; the nucleotide sequence database from the European Molecular Biology Laboratory; and NBRF-PIR, the National Biomedical Research Foundation's protein sequence database (Brutlag & Kristofferson, 1988).

Creating a new culture requires both carrot and stick. Making the online programs available and easy to use was one thing. Getting all molecular biologists to use them was another. In order to doubly encourage molecular biologists to comply with the new procedure of submitting their data online, the major molecular biology journals agreed to require evidence that data had been so submitted before they would consider a manuscript for review. *Nucleic Acids Research* was the first journal to enforce this transition to electronic data submission (Brutlag & Kristofferson, 1988). With these new policies and networks in place, BIONET was able to reduce the time from submission to publication and dis-

tribution of new sequence data from two years to twenty-four hours. As noted above, just a few years earlier, at the beginning of BIONET, there were only ten million base pairs published, and these had been the result of several years' effort. The new electronic submission of data generated ten million base pairs a month (Douglas Brutlag, personal communication, 19 June 1999; "Nomination for Smithsonian-ComputerWorld Award," n.d.). Walter Gilbert may have angered some of his colleagues at the 1987 Los Alamos Workshop on Automation in Decoding the Human Genome when he stated that "Sequencing the human genome is not science, it is production" (Boswell, 1987). But he surely had his finger on the pulse of the new biology.

The Matrix of Biology

The explosion of data on all levels of the biological continuum made possible by the new biotechnologies and represented powerfully by organizations such as BIONET was a source of both exhilaration and anxiety. Of primary concern to many biologists was how best to organize this massive outpouring of data in a way that would lead to deeper theoretical insight, perhaps even a unified theoretical perspective for biology. The National Institutes of Health were among those most concerned about these issues, and they organized a series of workshops to consider the new perspectives emerging from recent developments. The meetings culminated in a report from a committee chaired by Harold Morowitz titled *Models for Biomedical Research: A New Perspective* (1985). The committee foresaw the emergence of a new theoretical biology "different from theoretical physics, which consists of a small number of postulates and the procedures and apparatus for deriving predictions from those postulates." The new biology was far more than just a collection of experimental observations. Rather it was a vast array of information gaining coherence through organization into a conceptual matrix (Morowitz, 1985, p. 21). A point in the history of biology had been reached where new generalizations and higher-order biological laws were being approached but obscured by the simple mass of data and volume of literature. To move toward this new theoretical biology, the committee proposed a multidimensional matrix of biological knowledge:

> That is the complete data base of published biological experiments structured by the laws, empirical generalizations, and physical foundations of biology and connected by all the interspecific transfers of information. The matrix includes but is more than the computerized data base of biological literature, since the search methods and key words used in gaining access to that base are themselves related to the generalizations and ideas about the structure of biological knowledge. (Morowitz, 1985, p. 65)

New disciplinary requirements were imposed on the biologist who wanted to interpret and use the matrix of biological knowledge:

> The development of the matrix and the extraction of biological generalizations from it are going to require a new kind of scientist, a person familiar enough with the subject being studied to read the literature critically, yet expert enough in information science to be innovative in developing methods of classification and search. This implies the development of a new kind of theory geared explicitly to biology with its particular theory structure. It will be tied to the use of computers, which will be required to deal with the vast amount and complexity of the information, but it will be designed to search for general laws and structures that will make general biology much more easily accessible to the biomedical scientist. (Morowitz, 1985, p. 67)

Similar concerns about managing the explosion of new information motivated the Board of Regents of the National Library of Medicine. In its Long Range Plan of 1987 the NLM drew directly on the notion of the matrix of biological knowledge and elaborated upon it explicitly in terms of fashioning the new biology as an information science (Board of Regents, 1987). The Long Range Plan contained a series of recommendations that were the outcome of studies done by five different panels, including a panel that considered issues connected with building factual databases, such as sequence databases.

In the view of the panel the field of molecular biology was opening the door to an era of unprecedented understanding and control of life processes, including "automated methods now available to analyze and modify biologically important macromolecules" (Board of Regents, 1987, p. 26). The report characterized biomedical databases as representing the universal hierarchy of biological nature: cells, chromosomes, genes, proteins. Factual databases were being developed at all levels of the hierarchy, from cells to base-pair sequences. Because of the complexity of biological systems, basic research

in the life sciences was increasingly dependent on automated tools to store and manipulate the large bodies of data describing the structure and function of important macromolecules. The NIH Long Range Plan stated, however, that the critical questions being asked could often only be answered by relating one biological level to another, but methods for automatically suggesting links across levels were nonexistent (Board of Regents, 1987, pp. 26–27).

> A singular and immediate window of opportunity exists for the Library in the area of molecular biology information. Because of new automated laboratory methods, genetic and biochemical data are accumulating far faster than they can be assimilated into the scientific literature. The problems of scientific research in biotechnology are increasingly problems of information science. By applying its expertise in computer technologies to the work of understanding the structure and function of living cells on a molecular level, NLM can assist and hasten the Nation's entry into a remarkable new age of knowledge in the biological sciences. (Board of Regents, 1987, p. 29)

To support and promote the entry into the new age of biological knowledge, the NIH recommended building a National Center for Biotechnology Information to serve as a repository and distribution center for this growing body of knowledge and as a laboratory for developing new information analysis and communications tools essential to the advance of the field. The proposal recommended $12.75 million per year for 1988–1990, with an additional $10 million per year for work in medical informatics (Board of Regents, 1987, pp. 46–47). The program would emphasize collaboration between computer and information scientists and biomedical researchers. In addition the NIH would support research in the areas of molecular biology database representation, retrieval-linkages, and modeling systems, while examining interfaces based on algorithms, graphics, and expert systems. The recommendation also called for the construction of online data delivery through linked regional centers and distributed database subsets.

Brave New Theory

Two different styles of work have characterized the field of molecular biology. The biophysical approach has sought to predict the function of a molecule from its structure. The biochemical approach, on the other hand, has been concerned with predicting phenotype from biochemical function. If there has been a unifying framework for the field, at least from its early days up through the 1980s, it was provided by the "central dogma" emerging from the work of James Watson, Francis Crick, Monod, and Jacob in the late 1960s, schematized as follows:

DNA → RNA → Protein → Function

In this paper I have singled out molecular biologists whose Holy Grail has always been to construct a mathematized, predictive biological theory. In terms of the "central dogma" the measure of success in the enterprise of making biology predictive would be—and has been since the days of Claude Bernard—rational medicine. If one had a complete grasp of all the levels from DNA to behavioral function, including the processes of translation at each level, then one could target specific proteins or biochemical processes that may be malfunctioning and design drugs specifically to repair these disorders. For those molecular biologists with high theory ambitions, the preferred path toward achieving this goal has been based on the notion that the function of a molecule is determined by its three-dimensional folding and that the structure of proteins is uniquely contained in the linear sequence of their amino acids (Anfinsen, 1973). But determination of protein structure and function is only part of the problem confronting a theoretical biology. A fully fledged theoretical biology would want to be able to determine the biochemical function of the protein structure as well as its expected behavioral contribution within the organism. Thus biochemists have resisted the road of high theory and have pursued a solidly experimental approach aimed at eliciting common models of biochemical function across a range of mid-level biological structures from proteins and enzymes through cells. Their approach has been to identify a gene by some direct experimental procedure—determined by some property of its product or otherwise related to its phenotype—to clone it, to sequence it, to make its product, and to continue to work experimentally so as to seek an understanding of its function. This model, as Walter Gilbert has observed, was suited to "small science," experimental science conducted in a single lab (Gilbert, 1991, p. 99).

The emergence of organizations like the Brookhaven Protein Data Bank in 1971, GenBank in 1982, and BIONET in 1984, and the massive amount of sequencing data that began to become available in university and company databases, and more recently publicly through the Human Genome Initiative, has complicated this picture immensely through an unprecedented influx of new data. In the process a paradigm shift has occurred in both the intellectual and institutional structures of

biology. According to some of the central players in this transformation, at the core is biology's switch from having been an observational science, limited primarily by the ability to make observations, to being a data-bound science limited by its practitioners' ability to understand large amounts of *information* derived from observations. To understand the data, the tools of information science have not only become necessary handmaidens to theory; they have also fundamentally changed the picture of biological theory itself. A new picture of theory radically different from even the biophysicists' model of theory has come into view. In terms of discipline biology has become an information science. Institutionally, it is becoming "Big Science." Gilbert characterizes the situation sharply:

> To use this flood of knowledge, which will pour across the computer networks of the world, biologists not only must become computer-literate, but also change their approach to the problem of understanding life.
>
> The next tenfold increase in the amount of information in the databases will divide the world into haves and have-nots, unless each of us connects to that information and learns how to sift through it for the parts we need. (Gilbert, 1991)

The new data-bound biology implied in Gilbert's scenario is genomics. The theoretical component of genomics might be termed *computational biology*, while its instrumental and experimental component might be considered *bioinformatics*. The fundamental dogma of this new biology, as characterized by Douglas Brutlag, reformulates the central dogma of Jacob-Monod in terms of "information flow" (Brutlag, 1994):

Genetic information → Molecular structure → Biochemical function → Biologic behavior

Walter Gilbert describes the newly forming genomic view of biology:

> The new paradigm now emerging is that all the "genes" will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis. The actual biology will continue to be done as "small science"—depending on individual insight and inspiration to produce new knowledge—but the reagents that the scientist uses will include a knowledge of the primary sequence of the organism, together with a list of all previous deductions from that sequence. (Gilbert, 1991, p. 99)

Genomics, computational biology, and bioinformatics restructure the playing field of biology, bringing a substantially modified toolkit to the repertoire of molecular biology skills developed in the 1970s. Along with the biochemistry components, new skills are now required, including machine learning, robotics, databases, statistics and probability, artificial intelligence, information theory, algorithms, and graph theory (Douglas Brutlag, personal communication).

Proclamations of the sort made by Gilbert and other promoters of genomics may seem like hyperbole. But the Human Genome Initiative and the information technology that enables it have fundamentally changed molecular biology, and indeed, may suggest similar changes in store for other domains of science. The online DNA and protein databases that I have described have not just been repositories of information for insertion into the routine work of molecular biology, and the software programs discussed in connection with IntelliGenetics and GenBank are more than retrieval aids for transporting that information back to the lab. As a set of final reflections, I want to look in more detail at some ways this software has been used to address the problems of molecular biology in order to gain a sense of the changes taking place.

Biology in Silico

To appreciate the relationship between genomics and earlier work in molecular biology, it is useful to compare approaches to the determination of structure and function. Rather than an approach deriving structure and function from first principles of the dynamics of protein folding, the bioinformatics approach involves comparing new sequences with preexisting ones and discovering structure and function by homology to known structures. This approach examines the kinds of amino acid sequences or patterns of amino acids found in each of the known protein structures. The sequences of proteins whose structure have already been determined and are already on file in the PDP are examined to infer rules or patterns applicable to novel protein sequences to predict their structure. For instance, certain amino acids, such as leucine and alanine, are very common in α-helical regions of proteins, whereas other amino acids, such as proline, are rarely if ever found in α-helices. Using patterns of amino acids or rules based on these patterns, the genome scientist can attempt to predict

where helical regions will occur in proteins whose structure is unknown and for which a complete sequence exists. Clearly the lineage in this approach is work on automated learning first begun in DENDRAL and carried forward in other AI projects related to molecular biology such as MOLGEN.

The great challenge in the study of protein structure has been to predict the fold of a protein segment from its amino acid sequence. Before the advent of sequencing technology it was generally assumed that each unique protein sequence would produce a three-dimensional structure radically different from every other protein. But the new technology revealed that protein sequences are highly redundant: Only a small percentage of the total sequence is crucial to the structure and function of the protein. Moreover, while similar protein sequences generally indicate similarly folded conformations and functions, the converse does not hold. In some proteins, such as the nucleotide-binding proteins, the structural features encoding a common function are conserved, while primary sequence similarity is almost nonexistent (Rossman, Moras, & Olsen, 1974; Creighton, 1983; Birktoft & Banaszak, 1984). Methods that detect similarities solely at the primary sequence level turned out to have difficulty addressing functional associations in such sequences. A number of features often only implicit in the protein's linear or primary sequence of twenty possible amino acids turned out to be important in determining structure and function.

Such findings implied the need for more sophisticated techniques of searching than simply finding identical matches between sequences in order to elicit information about similarities between higher-ordered structures such as folds. One solution adopted early on by programs such as SEQ was to assume that if two DNA segments are evolutionarily related, their sequences will probably be related in structure and function. The related descendants are identifiable as homologues. For instance, there are more than 650 globin sequences (as in myoglobin or hemoglobin) in the protein sequence databases, all of them very similar in structure. These sequences are assumed to be related by evolutionary descent rather than having been created de novo. Many programs for searching sequence databases have been written, including an important early method written in 1970 by S. B. Needleman and C. D. Wunsch and incorporated into SEQ for aligning sequences based on homologies (Needleman & Wunsch, 1970). The method of homology depends upon assumptions related to the genetic events that could have occurred in the divergent (or convergent) evolution of proteins; namely, that homologous proteins are the result of gene duplication and subsequent mutations. If one assumes that after the duplication point mutations occur at a constant or variable rate, but randomly along the genes of the two proteins, then after a relatively short period of time the protein pairs will have nearly identical sequences. Later there will be gaps in the shared sets of base-pairs between the two proteins. Needleman and Wunsch determined the degree of homology between protein pairs by counting the number of non-identical pairs (amino acid replacements) in the homologous comparison and using this number as a measure of evolutionary distance between the amino acid sequences. A second approach was to count the minimum number of mutations represented by the non-identical pairs.

Another example of a key tool used in determining structure-function relationship is a search for sequences that correspond to small conserved regions of proteins, modular structures known as motifs. Since insertions and deletions (gaps) within a motif are not easily handled from a mathematical point of view, a more technical term, "alignment block," has been introduced that refers to conserved parts of multiple alignments containing no insertions or deletions (Bork & Gibson, 1996).

Several different kinds of motifs are related to secondary and tertiary structure. Protein scientists distinguish among four hierarchical levels of structure. Primary structure is the specific linear sequence of the twenty possible amino acids making up the building blocks of the protein. Secondary structure consists of patterns of repeating polypeptide structure within an α-helix, β-sheet, and reverse turns. Supersecondary structure refers to a few common motifs of interconnected elements of secondary structure. Segments of α-helix and β-strand often combine in specific structural motifs. One example is the α-helix-turn-helix motif found in DNA-binding proteins. This motif contains twenty-two amino acids in length that enable it to bind to DNA. Another motif at the supersecondary level is known as the Rossmann fold, in which three α-helices alternate with three parallel β-strands. This has turned out to be a general fold for binding mono- or dinucleotides and is the most common fold observed in globular proteins (Richardson & Richardson, 1989).

A higher order of modular structure is found at the tertiary level. Tertiary structure is the overall spatial arrangement of the polypeptide chain into a globular mass of hydrophobic side chains forming the central core, from which water is excluded, and more polar side chains fa-

voring the solvent-exposed surface. Within tertiary structures are certain domains on the order of a hundred amino acids, which are themselves structural motifs. Domain motifs have been shown to be encoded by exons, individual DNA sequences that are directly translated into peptide sequences. Assuming that all contemporary proteins have been derived from a small number of original ones, Walter Gilbert and colleagues have argued that the total number of exons from which all existing protein domains have been derived is somewhere between one thousand and seven thousand (Dorit, Schoenback, & Gilbert, 1990).

Motifs are powerful tools for searching databases of known structure and function to determine the structure and function of an unknown gene or protein. The motif can serve as a kind of probe for searching the database or some new sequence, testing for the presence of that motif. The PROCITE database, for example, has more than a thousand of these motifs (Bairoch, 1991). With such a library of motifs one can take a new sequence and use each one of the motifs to get clues about its structure. Suppose, for example, the sequence of a gene or protein has been determined. Then the most common way to investigate its biologic function is simply to compare its sequence with all known DNA or protein sequences in the databases and note any strong similarities. The particular gene or protein that has just been determined will of course not be found in the databases, but a homologue from another organism or a gene or protein having a related function may be found. The evolutionary similarity implies a common ancestor and hence a common function. Searching with motif probes refines the determination of the fold regions of the protein. These methods become more and more successful as the databases grow larger and as the sensitivity of the search procedure increases. Bork, Ouzounis, and Sander (1994) state that the likelihood of identifying homologues is currently higher than 80 percent for bacteria, 70 percent for yeast, and about 60 percent for animal sequence series (Bork & Gibson, 1996).

The all-or-nothing character of consensus sequences—a sequence either matches or it does not—led researchers to modify this technique to introduce degrees of similarity among aligned sequences as a way of detecting similarities between proteins, even distantly related ones. Knowing the function of a protein in some genome, such as *E. coli*, for instance, might suggest the same function of a closely related protein in an animal or human genome (Patthy, 1996). Moreover, as noted above, different amino acids can fit the same pattern, such as the helix-turn-helix, so that a representation of sequence pattern in which alternative amino acids are acceptable, as well as regions in which a variable number of amino acids may occur, are desirable ways of extending the power of straightforward consensus sequence comparison. One such technique is to use weights or frequencies to specify greater tolerance in some positions than in others. An illustration of the success of this approach is provided by the DNA-binding proteins mentioned above, which contain a helix-turn-helix motif twenty-two acids in length (Brennan & Mathews, 1989). Comparison of the linear amino acid sequences of these proteins revealed no consensus sequence that could distinguish them from any other protein. A weight matrix is constructed by determining the frequency with which each amino acid appears at each position, and then converting these numbers to a measure of the probability of occurrence of each acid. This weight matrix can be applied to measure the likelihood that any given sequence twenty-two amino acids long is related to the helix-turn-helix family. A further modification of the weight matrix is the profile, which allows one to estimate the probability that any amino acid will appear in a specific position (Gribskov et al., 1987; Gribskov et al., 1988).

In addition to consensus sequences, weight matrices, and profiles, a further class of strategies for determining structure-function relations are various sequence alignment methods. In order to detect homologies between distantly related proteins, one method is to assign a measure of similarity to each pair of amino acids, and then add up these pairwise scores for the entire alignment (Schwartz & Dayhoff, 1979). Related proteins will not have identical amino acids aligned, but they do have chemically similar or replaceable amino acids in similar positions. In a scoring method developed by R. M. Schwartz and M. O. Dayhoff, for example, amino acid pairs that are identical or chemically similar were given positive scores, and pairs of amino acids that are not related were assigned negative similarity scores.

A dramatic illustration of how sequence alignment tools can be brought to bear on determining function and structure is provided by the case of cystic fibrosis. Cystic fibrosis is caused by aberrant regulation of chloride transport across epithelial cells in the pulmonary tree, the intestine, the exocrine pancreas, and apocrine sweat glands. This disorder was identified as being caused by defects in the cystic fibrosis transmembrane conductance regulator protein (CFTR). After the CFTR gene was isolated in 1989, its protein product was identified as producing a chloride channel, which depends for

its activity on the phosphorylation of particular residues within the regulatory region of the protein. Using computer-based sequence alignment tools of the sort described above, it was established that a consensus sequence for nucleotide binding folds that bind ATP are present near the regulatory region and that 70 percent of cystic fibrosis mutations are accounted for by a three base-pair deletion that removes a phenylalanine residue within the first nucleotide-binding position. A significant portion of the remainder of cystic fibrosis mutations affect a second nucleotide-binding domain near the regulatory region (Hyde et al., 1990; Kerem et al., 1989; Kerem et al., 1990; Riordan et al., 1989).

In working out the folds and binding domains for the CFTR protein, S. C. Hyde, P. Emsley, M. J. Hartshorn, and colleagues (1990) used sequence alignment methods similar to those available in early models of the IntelliGenetics software suite. They used the Chou-Fasman algorithm (1973) for identifying consensus sequences and the Quanta modeling package produced by Polygen Corporation (Waltham, Massachusetts) for modeling the protein and its binding sites (Hyde et al., 1990). In 1992 IntelliGenetics introduced BLAZE, an even more rapid search program running on a massively parallel computer. As an example of how computational genomics can be used to solve structure-function problems in molecular biology, Brutlag repeated the CFTR case using BLAZE (Brutlag, 1994). A sequence similarity search compared the CFTR protein to more than twenty-six thousand proteins in a protein database of more than nine million residues, resulting in a list of twenty-seven top similar proteins, all of which strongly suggested the CFTR protein is a membrane protein involved in secretion. Another feature of the comparison result was that significant homologies were shown with ATP-binding transport proteins, further strengthening the identification of CFTR as a membrane protein. The search algorithm identified two consensus sequence motifs in the protein sequence of the cystic fibrosis gene product that corresponded to the two sites on the protein involved in binding nucleotides. The search also turned up distant homologies between the CFTR protein and proteins in *E. coli* and yeast. The entire search took three hours. Such examples offer convincing evidence that tools of computational molecular biology can lead to the understanding of protein function.

The methods for analyzing sequence data discussed above were just the beginnings of an explosion of database mining tools for genomics that is continuing to take place.[3] In the process biology is becoming even more aptly characterized as an information science (Hughes et al., 1999; IntelliGenetics & MasPar Computer Corporation, 1992). Advances in the field have led to large-scale automation of sequencing in genome centers employing robots. The success this large-scale sequencing of genes has enjoyed has in turn spawned a similar approach to applying automation to sequencing proteins, a new area complementary to genomics called proteomics. Similar in concept to genomics, which seeks to identify all genes, proteomics aims to develop techniques that can rapidly identify the type, amount, and activities of the thousands of proteins in a cell. Indeed, new biotechnology companies have started marketing technologies and services for mining protein information en masse. Oxford Glycosciences (OGS) in Abingdon, England, has automated the laborious technique of two-dimensional gel electrophoresis.[4] In the OGS process an electric current applied to a sample on a polymer gel separates the proteins, first by their unique electric charge characteristics and then by size. A dye attaches to each separated protein arrayed across the gel. Then a digital imaging device automatically detects protein levels by how much the dye fluoresces. Each of the five thousand to six thousand proteins that may be assayed in a sample in the course of a few days is channeled through a mass spectrometer that determines its amino acid sequence. The identity of a protein can be determined by comparing the amino acid sequence with information contained in numerous gene and protein databases. One imaged array of proteins can be contrasted with another to find proteins specific to a disease.

In order to keep pace with this flood of data emerging from automated sequencing, genome researchers have in turn looked increasingly to artificial intelligence, machine learning, and even robotics in developing automated methods for discovering patterns and protein motifs from sequence data. The power of these methods is their ability both to represent structural features rather than strictly evolutionary steps and to discover motifs from sequences automatically. The methods developed in the field of machine learning have been used to extract conserved residues, discover pairs of correlated resi-

[3] See, for instance, the National Institute of General Medical Science (NIGMS) "Protein Structure Initiative Meeting Summary," 24 April 1998 at http://www.nih.gov/nigms/news/reports/protein_structure.html.

[4] See the discussion of this technology at the site of Oxford Glycosciences: http://www.ogs.com/proteome/home.html.

dues, and find higher-order relationships between residues as well. Techniques from the field of machine learning have included perceptrons, discriminant analysis, neural networks, Bayesian networks, hidden Markov models, minimal length encoding, and context-free grammars (Hunter, 1993). Important methods for evaluating and validating novel protein motifs have also derived from the machine learning area.

An example of this effort to scale up and automate the discovery of structure and function is EMOTIF (for "electronic-motif"), a program for discovering conserved sequence motifs from families of aligned protein sequences developed by the Brutlag Bioinformatics Group at Stanford (Nevill-Manning et al., 1998).[5] Protein sequence motifs are usually generated manually with a single "best" motif optimized at one level of specificity and sensitivity. Brutlag's aim was to automate this procedure. An automated method requires knowledge about sequence conservation. For EMOTIF, this knowledge is encoded as a particular allowed set of amino acid substitution groups. Given an aligned set of protein sequences, EMOTIF works by generating a set of motifs with a wide range of specificities and sensitivities. EMOTIF can also generate motifs that describe possible subfamilies of a protein superfamily. The EMOTIF program works by generating a new database, called IDENTIFY, of fifty thousand motifs from the combined seven thousand protein alignments in two widely used public databases, the PRINTS and BLOCKS databases. By changing the set of substitution groups, the algorithm can be adapted for generating entirely new sets of motifs.

Highly specific motifs are well suited for searching entire proteomes. IDENTIFY assigns biological functions to proteins based on links between each motif and the BLOCKS or PRINTS databases that describe the family of proteins from which it was derived. Because these protein families typically have several members, a match to a motif may provide an association with several other members of the family. In addition, when a match to a motif is obtained, that motif may be used to search sequence databases, such as SWISS-PROT and GenPept, for other proteins that share this motif. In their paper introducing these new programs C. G. Nevill-Manning, T. D. Wu, and Brutlag showed that EMOTIF and IDENTIFY successfully assigned functions automatically to 25 to 30 percent of the proteins in several bacterial genomes and automatically assigned functions to 172 proteins of previously unknown function in the yeast genome.

Many molecular biologists who welcomed the Human Genome Initiative with open arms undoubtedly believed that when the genome was sequenced everyone would return to the lab to conduct their experiments in a business-as-usual fashion, empowered with a richer set of fundamental data. The developments in automation, the resulting explosion of data, and the introduction of tools of information science to master this data have changed the playing field forever: There may be no "lab" to return to. In its place is a workstation hooked to a massively parallel computer, producing simulations by drawing on the data streams of the major databanks, and carrying out "experiments" *in silico* rather than in vitro. The result of biology's metamorphosis into an information science just may be the relocation of the lab to the industrial park and the dustbin of history.

References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science, 181*(4096), 223–230.

Bairoch, A. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Research, 19,* 2241.

Bernstein, F. C., Koetzle, T. F., et al. (1977). The Protein Data Bank: A computer based archival file for macromolecular structure. *Journal of Molecular Biology, 112,* 535–542.

BIONET users status. (1986, April 3 and October 9). From BIONET managers' meetings. (Stanford University Special Collections, Brutlag Papers, Fldr BIONET). Stanford, CA: Stanford University.

Birktoft, J. J., & Banaszak, L. J. (1984). Structure-function relationships among nicotinamide-adenine dinucleotide dependent oxidoreductases. In M. T. W. Hearn (Ed.), *Peptide and Protein Reviews* (Vol. 4, pp. 1–47). New York: Marcel Dekker.

Board of Regents. (1987). *NLM long range plan (report of the Board of Regents).* Bethesda, MD: National Library of Medicine.

Bork, P., & Gibson, T. J. (1996). Applying motif and profile searches. In R. F. Doolittle (Ed.), *Computer Methods for Macromolecular Sequence Analysis* (Vol. 266 in Methods in Enzymology, pp. 162–184, especially p. 163). San Diego: Academic Press.

Bork, P., Ouzounis, C., & Sander, C. (1994). From genome sequences to protein function. *Current Opinions in Structural Biology, 4*(3), 393–403.

Boswell, S. (January 9, 1987). Los Alamos workshop—Exploring the role of robotics and automation in decoding the human genome. (IntelliGenetics trip report. In Stanford Special Collections, Brutlag Papers, Fldr BIONET). Stanford, CA: Stanford University.

Brennan, R. G., & Mathews, B. W. (1989). The helix-turn-helix binding motif. *Journal of Biological Chemistry, 264,* 1903.

Brutlag, D. L. (1994). Understanding the human genome. In P. Leder, D. A. Clayton, & E. Rubenstein (Eds.), Scientific American: *Introduction to Molecular Medicine* (pp. 153–168). New York: Scientific American, Inc.

[5] EMOTIF can be viewed at http://motif.stanford.edu/emotif.

Brutlag, D. L., & Kristofferson, D. (1988). BIONET: An NIH computer resource for molecular biology. In R. R. Colwell (Ed.), *Biomolecular data: A resource in transition* (pp. 287–294). Oxford: Oxford University Press.

Business plan for IntelliGenetics. (1981, May 8). (Stanford Special Collections, Brutlag Papers, Fldr IntelliGenetics, p. 2). Stanford, CA: Stanford University.

Creighton, T. E. (1983). *Proteins: Structure and molecular properties.* New York: W. H. Freeman.

Dorit, R. L., Schoenback, L., & Gilbert, W. (1990). How big is the universe of exons? *Science, 250,* 1377.

Doyle, R. (1997). *On beyond living: Rhetorical transformations of the life sciences.* Stanford, CA: Stanford University Press.

Duke, B. H., Contracting Officer, NIH, to IntelliGenetics, Inc. (1987, June 3). Request for revised proposal in response to request for proposals RFP no. NIH-GM-87-04 titled "Nucleic Acid Sequence Data Bank." Letter with attachment. (Stanford Special Collections, Brutlag Papers, Fldr BIONET). Stanford, CA: Stanford University.

Feigenbaum, E. A., Buchanan, B., et al. (April 1980). A proposal for continuation of the MOLGEN project: A computer science application to molecular biology (Computer Science Department, Stanford University, Heuristic Programming Project, Technical Report No. HPP-80-5, Section 1), p. 1. Stanford, CA: Stanford University.

Feigenbaum, E. A., & Martin, N. (1977). Proposal: MOLGEN—a computer science application to molecular genetics (Heuristic Programming Project, Stanford University, Technical Report No: HPP-78-18,1977). Stanford, CA: Stanford University.

Friedhoff, R. M., & Benzon, W. (1989) *The second computer revolution: Visualization.* New York: W. H. Freeman.

Friedland, P. (1979). *Knowledge-based experiment design in molecular genetics.* Unpublished doctoral dissertation, Stanford University.

Friedland, P. (1984, April 27). BIONET organizational plans. Company confidential memo. (Stanford University Special Collections, Brutlag Papers, Fldr BIONET, p. 1). Stanford, CA: Stanford Universisty.

Friedland, P., Brutlag, D. L., Clayton, J., & Kedes, L. H. (1982). SEQ: A nucleotide sequence analysis and recombinant system. *Nucleic Acids Research, 10,* 279–294.

Friedland, P., to Reimers, N. Subject: Software licensing agreement: April 2, 1984. (Stanford University Special Collections, Fldr IntelliGenetics). Stanford, CA: Stanford University.

Gilbert, W. (1991). Towards a paradigm shift in biology. *Nature, 349,* 99.

Gribskov, M., Homyak, M., et al. (1988). Profile scanning for three-dimensional structural patterns in protein sequences. *Computer Applications in the Biosciences, 4,* 61.

Gribskov, M., McLachlan, A. D., et al. (1987). Profile analysis: Detection of distantly related proteins. *Proceedings of the National Academy of Sciences, 84,* 4355.

Hall, S. S. (1995). Protein images update natural history. *Science, 267*(3 February), 620–624.

Holbrook, S. R., Muskal, S. M., et al. (1993). Predicting protein structural features with artificial neural networks. In L. Hunter (Ed.), *Artificial intelligence and molecular biology* (pp. 161–194). Menlo Park, CA: AAAI (American Association for Artificial Intelligence) Press.

Huberman, J. (1989). BIONET: Computer power for the rest of us. (Stanford Special Collections, Brutlag Papers, Fldr BIONET). Stanford, CA: Stanford University.

Hughes, T. P., et al. (Eds.). (1999). *Funding a revolution: Government support for computing research.* Washington, DC: National Academy Press.

Hunter, L. (Ed.). (1993). *Artificial intelligence and molecular biology.* Menlo Park, CA: AAAI Press.

Hyde, S. C., Emsley, P., et al. (1990). Structural model of ATP-binding proteins associated with cystic fibrosis, multidrug resistance and bacterial transport. *Nature, 346,* 362–365.

IntelliGenetics. (1987). *Introduction to BIONET: A computer resource for molecular biology. User manual for Bionet subscribers, Release 2.3.* Mountain View, CA: IntelliGenetics.

Katz, L., & Levinthal, C. (1972). Interactive computer graphics and the representation of complex biological structures. *Annual Reviews in Biophysics and Bioengineering, 1,* 465–504.

Kerem, B. S., Rommens, J. M., et al. (1989). Identification of the cystic fibrosis gene: Genetic analysis. *Science, 245,* 1073–1080.

Kerem, B. S., Zielenski, J., et al. (1990). Identification of mutations in regions corresponding to the two putative nucleotide (ATP)-binding folds of the cystic fibrosis gene. *Proceedings of the National Academy of Sciences, 87,* 8447–8451.

Korn, L. J., Queen, C. L., & Wegman, M. N. (1977). Computer analysis of nucleic acid regulatory sequences. *Proceedings of the National Academy of Sciences, 74,* 4516–4520.

Lederberg, J. (n.d.). How DENDRAL was conceived and born (Stanford Technical Reports 048087-54, Knowledge Systems Laboratory Report No. KSL 87-54). Stanford, CA: Stanford University.

Lederberg, J. (1969). Topology of molecules. In National Research Council Committee on Support of Research in the Mathematical Sciences (Ed.), *The mathematical sciences: A collection of essays* (pp. 37–51). Cambridge, MA: MIT Press.

Lederberg, J., Sutherland, G. L., Buchanan, B. G., & Feigenbaum, E. (1969 November). A heuristic program for solving a scientific inference problem: Summary of motivation and implementation (Stanford Technical Reports 026104, Stanford Artificial Intelligence Project Memo AIM-104). Stanford, CA: Stanford University.

Ledley, R. S. (1965). *Use of computers in biology and medicine.* New York: McGraw-Hill.

Levinthal, C. (1966). Molecular model-building by computer. *Scientific American, 214*(6), 42–52.

Levinthal, C., Barry, C. D., Ward, S. A., & Zwick, M. (1968). Computer graphics in macromolecular chemistry. In D. Secrest & J. Nievergelt (Eds.), *Emerging concepts in computer graphics* (pp. 231–253). New York: W. A. Benjamin.

Lewin, R. (1984). National networks for molecular biologists. *Science, 223,* 1379–1380.

Maxam, A. M., to GENET Community. (1982, August 23). Subject: Closing of GENET. (Stanford University Special Collections, Peter Friedland Papers, Fldr GENET). Stanford, CA: Stanford University.

McCormick, B. H., DeFanti, T. A., & Brown, M. D. (1987). Visualization in Scientific Computing. NSF Report. *Computer Graphics, 21*(6), special issue.

Minutes of the meeting of the National Advisory Committee for BIONET. (1985, March 23). (Final version prepared 1985, August 1). (Stanford University Special Collections, Brutlag Papers, Fldr BIONET, p. 4). Stanford, CA: Stanford University.

Morowitz, H. (1985). *Models for biomedical research: A new perspective.* Washington, DC: National Academy of Sciences Press.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology, 48,* 443.

Nevill-Manning, C. G., Wu, T. D., et al. (1998). Highly specific protein

sequence motifs for genomic analysis. *Proceedings of the National Academy of Sciences, USA, 95*(11), 5865–5871.

NIH Special Study Section. (March 17–19, 1983). BIONET, national computer resource for molecular biology. (Stanford University Special Collections, Brutlag Papers, p. 2). Stanford, CA: Stanford University.

Nomination for Smithsonian-Computerworld Award. (n.d.). (Stanford Special Collections, Brutlag Papers, Fldr Smithsonian Computerworld Award). Stanford, CA: Stanford University.

Oettinger, A. G. (1966). The uses of computers in science. *Scientific American, 215*(3), 161–172.

Panel on Information Technology and the Conduct of Research, National Academy of Sciences. *Information technology and the conduct of research: The user's view.* (1989). Washington, DC: National Academy of Sciences.

Patthy, L. (1996). Consensus approaches in detection of distant homologies. In R. F. Doolittle (Ed.), *Computer methods for macromolecular sequence analysis* (Vol. 266 in Methods in Enzymology, pp. 184–198). San Diego: Academic Press.

PC/Gene: Microcomputer software for protein chemists and molecular biologists, user manual. (1986). Mountain View, CA: IntelliGenetics, pp. 99–120.

Ramsey, D. M. (Ed.). (1968). *Image processing in biological science.* Berkeley/Los Angeles: University of California Press.

Richardson, J. S., & Richardson, D. C. (1989). Principles and patterns of protein conformation. In G. D. Gasman (Ed.), *Prediction of protein structure and the principles of protein conformation.* New York: Plenum Press.

Rindfleisch, T., Friedland, P., & Clayton, J. (1981). The GENET guest service on SUMEX (SUMEX-AIM Report). (Stanford University Special Collections, Friedland Papers, Fldr GENET). Stanford, CA: Stanford University.

Riordan, J. R., Rommens, J. M., et al. (1989). Identification of the cystic fibrosis gene: Cloning and characterization of complementary RNA. *Science, 245,* 1066–1073.

Rossman, M. G., Moras, D., & Olsen, K. W. (1974). Chemical and biological evolution of a nucleotide-binding protein. *Nature, 250,* 194–199.

Schwartz, R. M., & Dayhoff, M. O. (1979). Matrices for detecting distant relationships. *Atlas of Protein Structure, 5*(Supplement 3), 353.

Second cut at interaction language and procedure. (n.d.). In Chemistry project. (Edward Feigenbaum Papers, Stanford Special Collections SC-340, Box 13). Stanford, CA: Stanford University.

Smith, D. H., Brutlag, D., Friedland, P., & Kedes, L. H. (1986). BIONET: National computer resource for molecular biology. *Nucleic Acids Research, 14,* 17–20.

Staden, R. (1977). Sequence data handling by computer. *Nucleic Acids Research, 4,* 4037–4051.

Staden, R. (1978). Further procedures for sequence analysis by computer. *Nucleic Acids Research, 5,* 1013–1015.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research, 6,* 2602–2610.

Stefik, M. (1977). Inferring DNA structures from segmentation data. *Artificial Intelligence, 11,* 85–114.

Watson, H. C. (1969). The stereochemistry of the protein myoglobin. *Progress in Stereochemistry, 4,* 299–333.